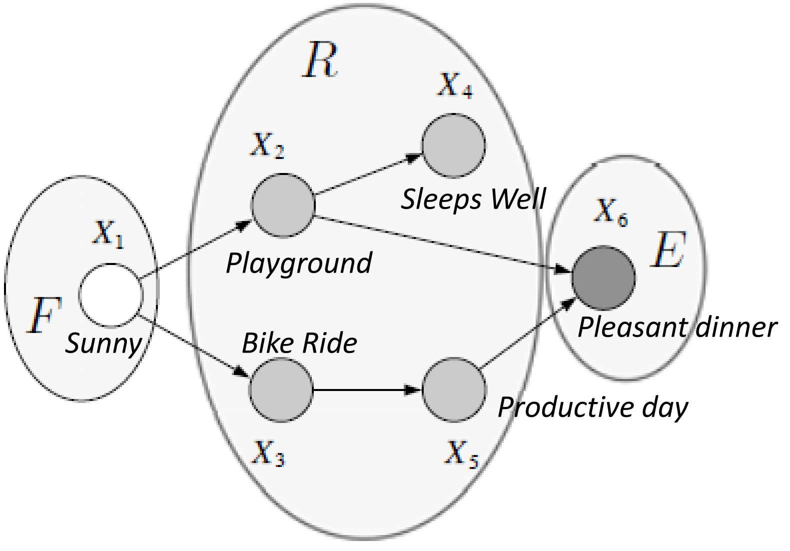
CORNELL
UNIVERSITY

Introduction to MCMC

DB Breakfast 09/30/2011 Guozhang Wang

Motivation: Statistical Inference



$$p(x_F|x_E) = \frac{\sum_{x_R} p(x_E, x_F, x_R)}{\sum_{x_R, x_F} p(x_E, x_F, x_R)}$$

Graphical Models

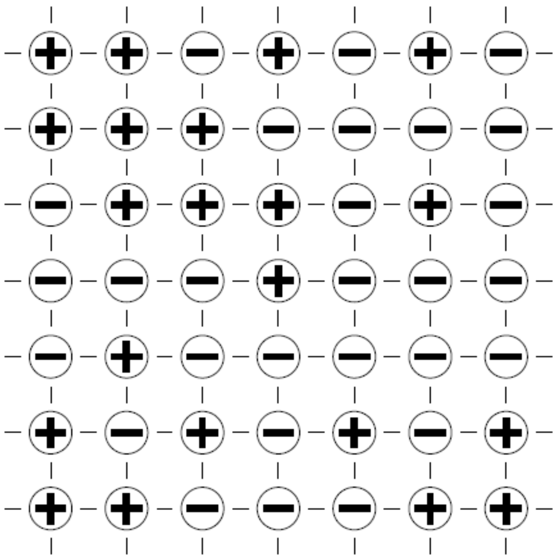
- Joint Distribution

$$p(x_1, \ldots, x_n) = \prod_v p(x_v | x_{\pi_v})$$

- Posterior Estimation

$$\mathrm{Var}(X) = \sum_{i=1}^{n} p_i \cdot (x_i - \mu)^2$$

Motivation: Statistical Physics



Ising Model

- Energy Model

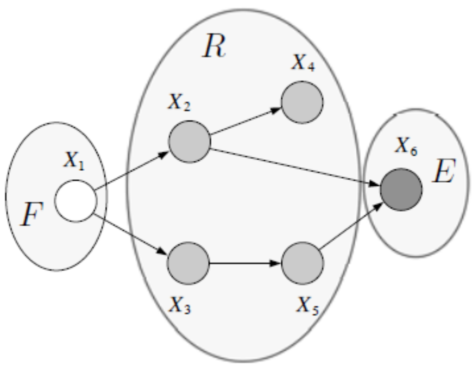
$$H(\sigma) = -\sum_{i \neq j} J_{ij}\sigma_i\sigma_j - \sum_j h_j\sigma_j$$

$$P(\sigma) = \frac{e^{-\beta H(\sigma)}}{Z}\,.$$

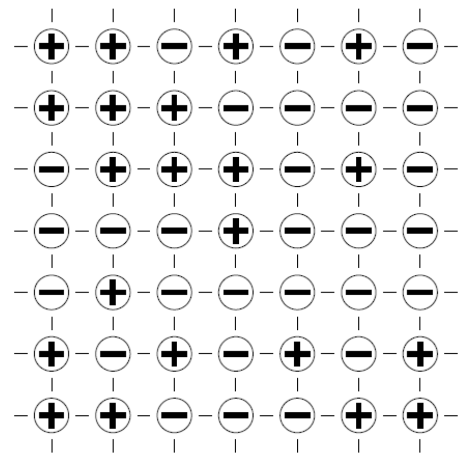
- Thermal Eqm. Estimation

$$E[\delta] = \sum_{i}^{2^n} \delta\, P(\delta)$$

Problem I: Integral Computation



$$p(x_F|x_E) = \frac{\sum_{x_R} p(x_E, x_F, x_R)}{\sum_{x_R, x_F} p(x_E, x_F, x_R)}$$



Posterior Estimation:

$$\mathrm{Var}(X) = \sum_{i=1}^{n} p_i \cdot (x_i - \mu)^2$$

Thermal Eqm. Estimation:

$$E[\delta] = \sum_{i}^{2^n} \delta \, P(\delta)$$

$$E[f(x)] = \int f(x)p(x)dx$$

Problem I Rewrite: Sampling

- Generate samples $\{x^{(r)}\}^R$ from the probability distribution p(x).

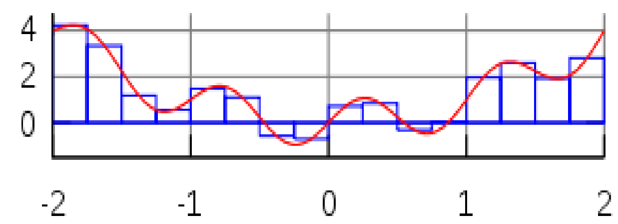
- If we can solve this problem, we can solve the integral computation by: $\sum_{i}^{R} f(x^{(r)})p(x^{(r)})$

- We will show later this estimator is ***unbiased*** with very nice ***variance bound***

Deterministic Methods

- Numerical Integration
  - Choose fixed points in the distribution
  - Use their probability values

$$\int_a^b f(x)\,dx \approx \frac{b-a}{n}\left(\frac{f(a)+f(b)}{2}+\sum_{k=1}^{n-1} f\left(a+k\frac{b-a}{n}\right)\right)$$



- Unbiased, but the variance is exponential to dimension

Random Methods: Monte Carlo

- Generate samples i.i.d
- Compute samples' probability
- Approximate integral by samples integration

$$\int f(x)p(x)dx \sim \sum f(X_i)p(X_i)$$

Merits of Monte Carlo

- Law of Large Numbers
  - Function f(x) over random variable x
  - **I.i.d** random samples drawn from p(x)

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \rightarrow \int f(x)p(x)dx \quad \textbf{\textit{as}} \quad n \longrightarrow \infty$$

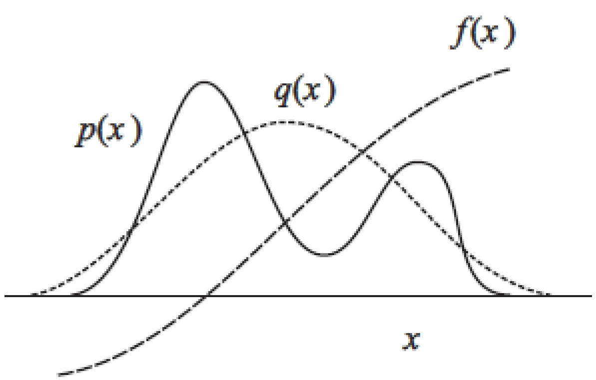
- Central Limit Theorem
  - **I.i.d** samples with expectation μ and variance σ^2

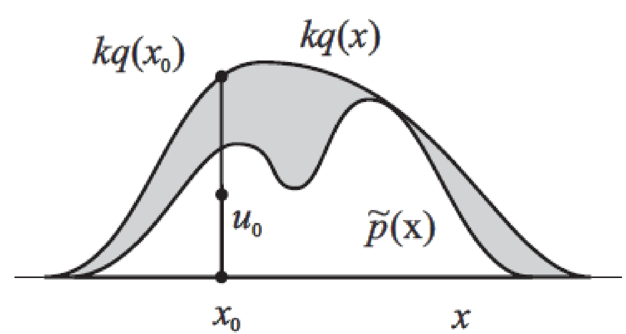
Sample distribution $\rightarrow$ normal(μ, σ^2/n)

Variance Not Depend on Dimension!

Simple Sampling

- Complex distributions
  - Known CDF: inversion methods
  - Simpler q(x) : Rejection sampling
  - Can compute density: importance sampling



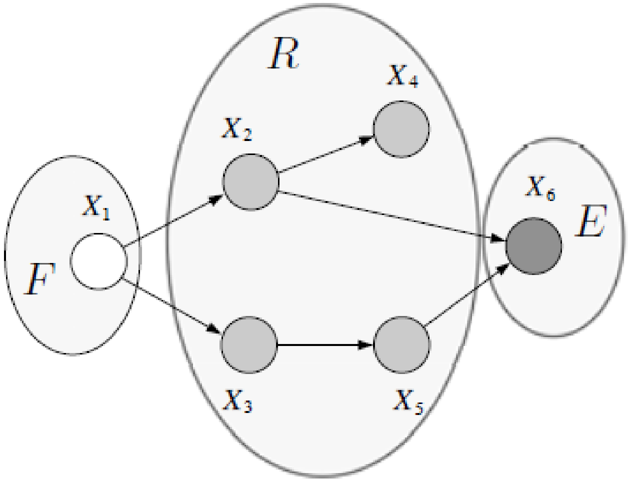


$$kq(x) \geq \tilde{p}(x) \ \forall x$$

Come Back to Statistical Inference

- Forward Sampling
  - Repeated sample $x_F^{(i)}$, $x_R^{(i)}$, $x_E^{(i)}$ based on prior and conditionals
  - Discard $x^{(i)}$ when $x_E^{(i)}$ is not observed x_E
  - When N samples retained, estimate $p(x_F|x_E)$ as

$$p(x_F|x_E) \approx \frac{1}{N}\sum_{i=1}^{N} I(x_F^{(i)} = x_F)$$

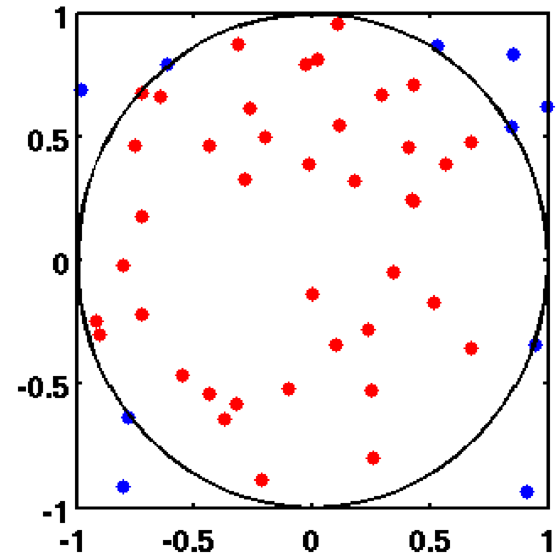


$$p(x_F|x_E) = \frac{\sum_{x_R} p(x_E, x_F, x_R)}{\sum_{x_R, x_F} p(x_E, x_F, x_R)}$$

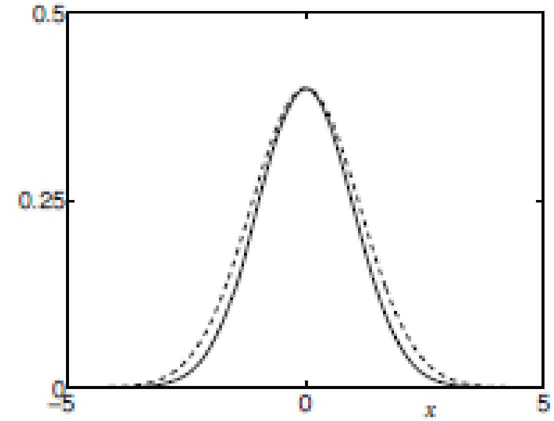
Problem: low acceptance rate

Problem II: Curse of Dimensionality

- The "prob. dense area" shrinks as dimension *d* arises



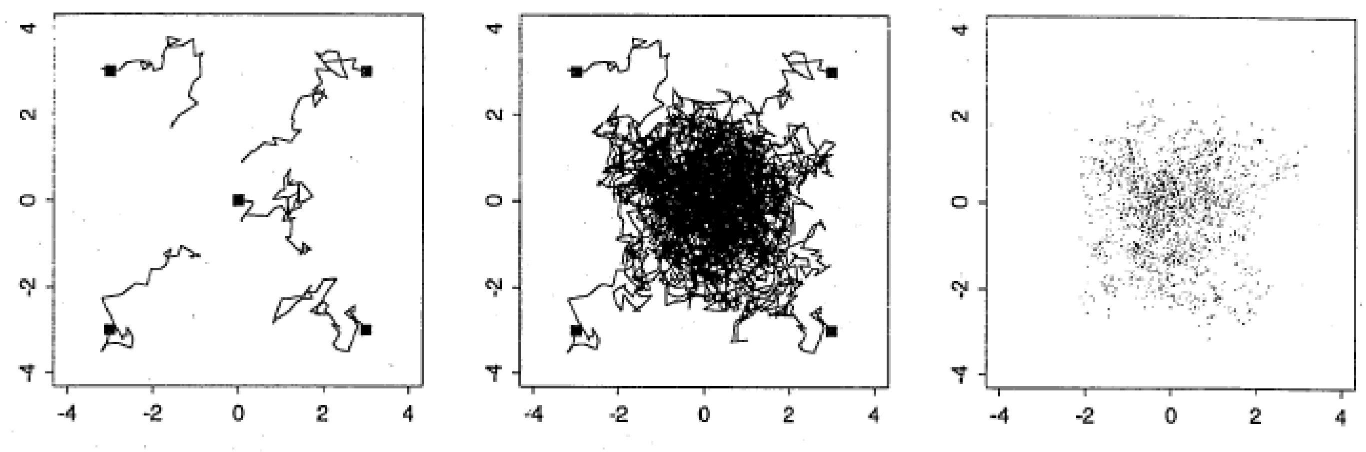
- Harder to sample in this area to get enough information of the distribution



- Acceptance rate decreases exponentially with *d*

Solution: Sampling with Guide

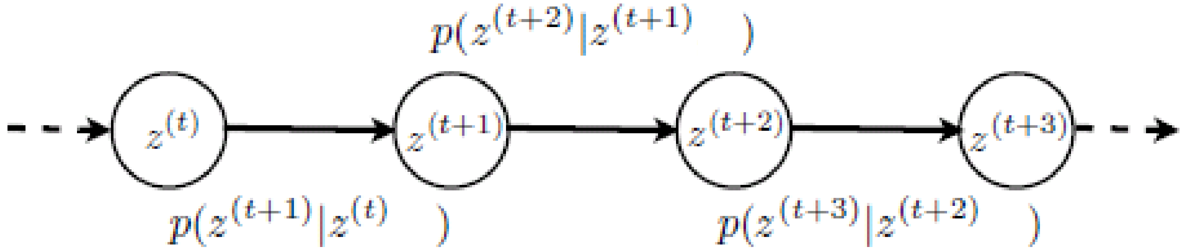
- Avoid random-walk, but sample variables conditional on previous samples



- Note: violate the i.i.d condition of LLN and CLT

Markov Chain

- Memoryless Random Process
  - Transition probability *A*: $p(x_{t+1}) = A * p(x_t)$



- Non-independent Samples, thus no guarantee of convergence

Mission Impossible?

How can we set the transition probabilities such that the 1) there is a equilibrium, and 2) equilibrium distribution is the target distribution, without knowing what the target is?

Markov Chain Properties

- A Markov chain is called:
  - *Stationary*, if there exists P such that P = A*P; note that multiple stationary distribution can exist.
  - *Aperiodic*, if there is no cycles with transition probability 1.
  - *Irreducible*, if has positive probability of reaching any state from any other
  - *Non-transient*, if it can always return to a state after visiting it
  - Reversible w.r.t *P*, if $P(x=i)\,A[ij] = P(x=j)\,A[ji]$

Convergence of Markov Chain

- If the chain is *Reversible* w.r.t. P, then P is its stationary distribution.
- And, if the chain is *Aperiodic* and *Irreducible,* it have a single stationary distribution, which it will converge to "almost surely".
- And, if the chain is *Non-transient*, it will always converge to its stationary distribution from any starting states.

Goal: Design alg. to satisfy all these properties.

Metropolis-Hastings

initialize with $z^{(0)}$ s.t. $p(z^{(0)} \mid x) > 0$

$t \leftarrow 1$

repeat

sample $z^{(t)}$ from $q(z^{(t)} \mid z^{(t-1)}, x)$

compute:

$$a(z^{(t-1)}, z^{(t)}) = \min\left(1, \frac{p(z^{(t)}|x)q(z^{(t-1)}|z^{(t)}, x)}{p(z^{(t-1)}|x)q(z^{(t)}|z^{(t-1)}, x)}\right)$$

draw u from $U(0,1)$

if $(u > a(z^{(t-1)}, z^{(t)}))$ $z^{(t)} \leftarrow z^{(t-1)}$ /* *reject proposal* */

if $(t > B$ and $t \bmod k = 0)$ retain sample $z^{(t)}$

$t \leftarrow t + 1$

until enough samples $(t = B + Sk)$

MCDB: A Monte Carlo Approach to Managing Uncertain Data

- Used for probabilistic Data management, where uncertainty can be expressed via distribution function.

```
CREATE TABLE SBP DATA(PID, GENDER, SBP) AS
 FOR EACH p in PATIENTS
  WITH SBP AS Normal (
   (SELECT s.MEAN, s.STD
    FROM SPB PARAM s))
  SELECT p.PID, p.GENDER, b.VALUE
  FROM SBP b
```

MCDB: A Monte Carlo Approach to Managing Uncertain Data

- Query processing
  - Sample instances from the distribution function
  - Execute the query on each sampled DB instance, thereby approximate the query-result distribution
  - Use Monte Carlo properties to compute mean, variance, quantiles, etc.
  - Some optimization Tricks
    - Tuple bundles
    - Split and merge

MCDB: A Monte Carlo Approach to Managing Uncertain Data

- Limits
  - Risk analysis concerns with quintiles mostly
  - Requires lots of samples to bound error
  - Actually is the curse of dimensionality
- MCDB-R: Risk Analysis in the Database
  - Monte Carlo + Markov Chain (MCMC)
  - Use Gibbs sampling

Thanks!